Ponencia

3

# LAS NUEVAS TECNOLOGÍAS EN LOS PROCESOS DE COOPERACIÓN DOCUMENTAL: AUMENTO DE LA VISIBILIDAD PARA REDINED

Sr. Laureano Felipe Gómez Dueñas. Docente e Investigador, Facultad Sistemas de Información y Documentación - Universidad de La Salle. Coordinador del Área de desarrollo tecnológico, Biblioteca Nacional de Colombia.

Índice

# LAS NUEVAS TECNOLOGÍAS EN LOS PROCESOS DE COOPERACIÓN DOCUMENTAL: AUMENTO DE LA VISIBILIDAD PARA REDINED

Sr. Laureano Felipe Gómez Dueñas. Docente e Investigador, Facultad Sistemas de Información y Documentación - Universidad de La Salle. Coordinador del Área de desarrollo tecnológico, Biblioteca Nacional de Colombia.

## Introducción

Este artículo busca hacer una aproximación al desarrollo de un modelo estándar de gestión documental digital desarrollada para REDINED, tomando este desarrollo como un medio para encontrar y ofrecer una solución a los cuatro problemas básicos en el manejo de información digital (Infoxicación, Acceso, Fiabilidad e Heterogenización de los sistemas de información). El modelo se basa en la implantación de estándares, protocolos y tecnologías documentales que buscan integrar y diseminar toda la información existente en la base de datos de REDINED, que utiliza un motor de base de datos desarrollado en CDS/ISIS. Esta información se encuentra actualmente invisible a los grandes sistemas de recuperación información (motores de búsqueda generales y específicos).

Se pretende llevar todo el contenido de la base de datos a su integración e inmersión total con los sistemas anteriormente mencionados, llevando a REDINED al nuevo paradigma de información denominado Internet 2.0. (Web semántica) y permitiendo la masificación de los contenidos existentes y unificando el sistema de información en el contexto universal de la recuperación de información.

## Los cuatro problemas principales al manejar información digital

Nos encontramos actualmente en una etapa de transición respecto al manejo de información, esto conlleva a generar un cambio de paradigmas para el manejo de los documentos digitales, vemos que hay un crecimiento desmesurado en la creación, diseminación y el uso de documentos digitales en las organizaciones e Internet, lo anterior nos conlleva a identificar cuatro problemas básicos que deben ser atacados

inmediatamente, estos son: la infoxicación, la fiabilidad, el acceso y la heterogenización de los sistemas de información documental.

**Infoxicación**

Podemos definir Infoxicación como el exceso de información circundante en nuestro entorno (documentos físicos, digitales, Internet...). Uno de los principales investigadores que trabajan el tema de la infoxicación, Alfons Cornella, nos expresa lo siguiente: "En cuanto al exceso de información, tenemos alguna duda de si efectivamente tenemos un exceso de información o somos nosotros los que no somos capaces de manejar toda la información que nos llega, a lo mejor nos quejamos por vicio, pero hay algo de realidad"[1]

Esta afirmación nos indica que todos somos conscientes de alguna manera del exceso de información que existe sobre Internet, pero aún no lo visualizamos como un problema. Por ejemplo nos gusta realizar búsquedas en Google y que este nos arroje más de 1.000 resultados, luego de esto pensamos: "que maravilla encontrar tantos resultados acerca de mi necesidad de información", sin embargo varios estudios demuestran que inconscientemente vamos preparados a revisar por mucho los primeros cincuenta resultados, esperando encontrar por lo menos cuatro documentos útiles.

Para visualizar el tamaño actual de la Web, podemos basarnos en un estudio realizado por la compañía inglesa Netcraft, compañía especializada en proveer servicios de seguridad en redes e Internet. Esta compañía publicó hace poco un informe sobre el crecimiento de Internet[2], en este informe indica que actualmente hay cerca de 101.435.253 sitios de Internet diferentes, y que en el último mes (Octubre 2006) se crearon cerca de 3.5 millones de sitios nuevos. Realizando cálculos muy simplistas, en el que cada sitio Web aporte como mínimo 50 documentos diferentes (páginas Web, documentos de Word, Excel, PDF entre otros), podríamos suponer que en Internet podemos encontrar cerca de 5.1 billones de documentos diferentes que cubren prácticamente cualquier área del conocimiento.

---

[1] Cornella, Alfons. Cómo sobrevivir a la infoxicación. Trascripción de la conferencia del acto de entrega de títulos de los programas de Formación de Posgrado del año académico 1999-2000. (Diciembre 2000)

[2] Netcraft. "November 2006 Web Server Survey". Web Server Survey News. http://news.netcraft.com/archives/2006/11/01/november_2006_web_server_survey.html. (Nov 1, 2006).

Después de ver estas cifras, es claro dilucidar que la producción intelectual se está generando a un ritmo mucho mayor de los que podemos nosotros ir consumiendo, de esta forma jamás podremos conocer el contenido de todas las páginas de Internet. Adicionalmente debido a la rápida evolución de las TICS y a la facilidad de uso y acceso a las mismas, nos vemos avocados a estar inundados de todo tipo de información digital y que esta continué creciendo exponencialmente, nos vemos sofocados ante tanta información y por lo tanto podemos afirmar que actualmente vivimos en una sociedad infoxicada.

**Fiabilidad**

Después de estimar la gran cantidad de documentos existentes en Internet, y retomando el problema anteriormente señalado, vemos que existe otro problema asociado: la dificultad de poder acceder a información pertinente y veraz. Sabemos que en Internet encontramos todo lo que pudiéramos imaginar y más, porque cualquier persona puede publicar información a su antojo sea esta veraz o no; de esta manera nos encontramos con el problema al determinar qué realmente me sirve de todo ese mar de documentos para solventar una necesidad de información, y aún más complicado, ¿como podría distinguir google entre información realizada con calidad de aquella que no es?.

"Uno de los problemas comunes en el análisis de contenidos de Internet es la carencia de procedimientos de análisis que certifiquen la fiabilidad de los documentos expuestos. La accesibilidad libre de los documentos a través de la Red obliga al usuario a determinar que su contenido cumple con determinadas expectativas de certeza y veracidad. No existe una verdad única y son las actitudes del usuario frente al documento las que definirán su valoración positiva o negativa"[3]. De acuerdo a lo anterior, vemos que la relevancia y pertinencia de un documento encontrado en Internet es muy subjetiva, sin embargo podemos observar con facilidad que, mucha información existente actualmente en Internet se puede considerar basura (incluye: paginas de productos y empresas comerciales, spam, etc..).

[3] Fornas Carrasco, Ricardo (2002) Criterios para evaluar la calidad y fiabilidad de los contenidos en Internet. In Proceedings Contenidos y Aspectos Legales en la Sociedad de la Información (CALSI), Valencia (España).

Un ejemplo significativo del problema en la fiabilidad de la información en Internet lo podemos evidenciar con el caso de la wikipedia, en el que uno de sus creadores Larry Sanger decidió abandonar este proyecto debido a varios incidentes e incongruencias en sus contenidos y crear uno nuevo llamado Citizendium, en el cual restringe la edición e ingreso de contenidos de acuerdo al nivel de reconocimiento del autor. El incidente más serio de la Wikipedia involucra un texto que acusaba a John Seigenthaler, ex periodista y creador del reconocido diario USA Today, de haber sido "patrocinador directo y autor intelectual de los asesinatos de John Kennedy y su hermano Bobby"[4]. Estos contenidos errados estuvieron varios meses al público antes de ser descubiertos, aún no se sabe cuánta gente llegó a leer el artículo difamatorio antes de haber sido corregido. "Y es que la causa de su mayor potencial es, a un mismo tiempo, la de su gran punto débil: al ser un proyecto totalmente abierto, en el que cualquiera puede añadir la información que desee. Los contenidos no dejan de expandirse, pero al mismo tiempo su fiabilidad ha sido puesta en repetidas ocasiones en tela de juicio por diferentes escándalos"[5].

**Acceso**

Aunque indiscutiblemente en Internet se encuentra información de calidad, acceder a esta información es difícil porque generalmente se encuentra oculta para el público en particular (**Internet invisible**), ó se requiere previamente estar inscrito en un servicio de información de pago (tanto para acceder al sistema como al documento en texto completo).

Para definir "**Internet Invisible"**, tomaremos el concepto de Bergam[6] que la define como el conjunto de todas las bases de datos y colecciones de documentos académicos relevantes que no son recuperables mediante el uso de un motor de búsqueda de propósito general como google. Generalmente estos documentos se encuentran indexados por buscadores científicos ó catálogos de bibliotecas sin que haya posibilidad de conectarlos a los buscadores generales. Como afirma Ricardo Baeza Yates en Excavando la Web*:* "la Web tiene actualmente al menos unas cuatro mil millones de páginas estáticas y un número cientos de veces mayor de dinámicas (aquellas que sólo

[4] Colaboradores de Wikipedia. Controversia por la biografía de John Seigenthaler Sr. [en línea]. Wikipedia, La enciclopedia libre, 2006 [fecha de consulta: 3 de noviembre del 2006]. Disponible en <http://es.wikipedia.org/w/index.php?title=Controversia_por_la_biograf%C3%ADa_de_John_Seigenthaler_Sr.&oldid=5411696>.

[5] FRANCIS PISANI Mejorar la Wikipedia y aprender a usarla. El Pais.Com. 05/01/2006. http://www.elpais.com/articulo/semana/Mejorar/Wikipedia/aprender/usarla/elpcibsem/20060105elpciblse_2/Tes/

[6] Bergman, M.K. (2001). The Deep Web: Surfacing Hidden Value. Journal of Electronic Publishing, 7(1).

se crean producto de un clic o de una consulta en un sitio Web). Además, tenemos que agregar toda la Web invisible (Web oculta) en intranets o páginas con acceso restringido. La Web invisible (Web oculta) es seguramente miles de veces más grande que la pública".

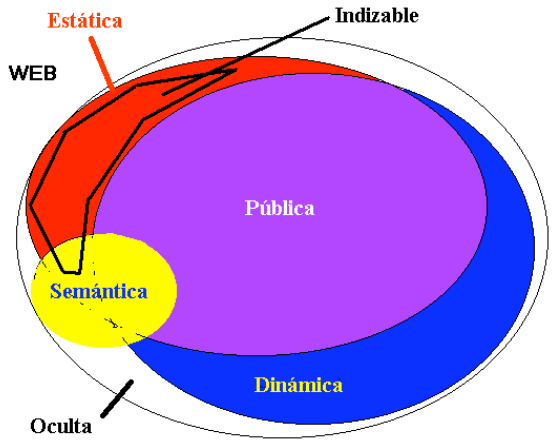


Figura 1. Tamaño de la Web [7]

Respecto al acceso de información de pago vemos que, los grandes proveedores de bases de datos cobran a las bibliotecas grandes sumas de dinero para acceder a sistemas de información documentales (para ellos acceder a información académica y de calidad es un derecho que hay que pagar). Por ejemplo muchos proveedores venden el acceso a una base de datos documental especializada en educación llamada Eric, sin embargo se puede acceder a esta base de datos de manera gratuita, ingresando directamente en el portal del ministerio de educación de los Estados Unidos (http://www.eric.ed.gov ).

**Heterogenización de los Sistemas de Información**

El problema de la Heterogenización está en la disparidad de mecanismos y procedimientos para el manejo de información digital que poseen los sistemas informáticos de gestión de documentos, así como de la forma de interactuar en Internet (Administradores de contenido-"CMS", Blogs, Wikis, Foros, Chats, Portales, Administradores documentales-"DMS", sistemas de información bibliográfica, etc..). Aún al interior de las mismas bibliotecas y organizaciones, encontramos que algunos de sus sistemas de información fueron diseñados generalmente para atender necesidades particulares, sin seguir ninguna norma o estándar. Debido a esto vemos una gran dificultad en que los sistemas de información se puedan interconectar y dialogar transparente y automáticamente sin la intervención humana que les permita compartir su

[7] Ricardo Baeza Yates. Excavando la Web. http://www.dcc.uchile.cl/~rbaeza/inf/webfaces.gif

información y documentos plenamente. "Existe un problema profundo a nivel interno de hetereogenización de productos de hardware y software, además de la forma de ingreso de la información (niveles de catalogación), ya que existen muchos datos repetidos, dispares y contradictorios si comparamos cada sistema existente"[8].

Pese a lo anterior, existe un gran interés en las bibliotecas y los productores de información académica en las universidades, para compartir su información mediante la integración de sus sistemas y catálogos bibliográficos con el ánimo de fortalecer sus colecciones y servicios de información, y esperando obtener mayor visibilidad e impacto principalmente en Internet, por ejemplo: "Los bibliotecólogos asumen que una copia de casi cada libro impreso reside en por lo menos una biblioteca o archivo en alguna parte en el mundo, y ellos sienten que como especialistas de información lo deben poder encontrar fácil y rápidamente"[9].

**Cómo solucionar los cuatro problemas**

Si planteamos que la unión de colecciones físicas existentes en una red de bibliotecas puede crear una mega biblioteca, ahora con el auge de las colecciones digitales que se están creando al interior de las organizaciones podemos afirmar rotundamente que, una perfecta unión y sincronización de todos los contenidos almacenados en los sistemas de información, permitiría a cualquier usuario del mundo acceder a **TODO** tipo de información académica sin ninguna limitante (social, económica, geográfica y política).

Así para solucionar los problemas anteriormente citados, desearíamos que google y cualquier buscador de información (especialmente académico), pudieran acceder e indexar transparentemente información de primera calidad (sistemas de información académicos y de investigación como REDINED), dándole mayor relevancia a estos contenidos y mostrándolos en las primeras posiciones cuando un usuario realice una búsqueda, aunque ello no solucionaría el problema de la infoxicación nos ahorraría tiempo porque sabríamos que siempre en los primeros resultados encontraremos información totalmente pertinente y de calidad.

---

[8] Gómez Dueñas, Laureano Felipe, La Iniciativa de Archivos Abiertos (OAI), un nuevo paradigma en la comunicación científica y el intercambio de información. En Revista Códice, Universidad de La Salle. No 4, 2005.

[9] COLE TIMOTHY W. Using OAI: innovations in the sharing of information. **En:** Library Hi Tech. Vol. 21 (2) . 2003. p. 115-117. (http://www.emeraldinsight.com/0737-8831.htm)

Para realizar ese intercambio y cooperación "transparente" entre varios sistemas de información, debemos hablar del término conocido como **interoperabilidad** (término a menudo traducido como interoperatibilidad, del inglés interoperability). Ésta podemos definirla como la capacidad de un sistema de información de comunicarse y compartir información efectivamente con otro mediante una interconexión libre y transparentemente (compartir documentos, metadatos y objetos digitales), sin dejar de utilizar en ningún momento la interfaz del sistema propio (siendo generalmente estos sistemas completamente heterogéneos, distribuidos y geográficamente distantes). "La interoperabilidad es una propiedad que puede predicarse de sistemas de naturaleza muy diferente, como pueden ser los sistemas informáticos, o los ferroviarios"[10].

Para lograr la interoperabilidad en los sistemas de información documentales, esta se debe manejar en distintos niveles: infraestructura, estructura, sintaxis y semántica[11].

- **Sintaxis**: Uso de herramientas e interfaces comunes que proporcionan uniformidad superficial en la navegación y el acceso.
- **Semántica**: Capacidad de los sistemas de información para acceder, de forma consistente y coherente, a objetos digitales y servicios similares, distribuidos en repositorios heterogéneos, con la ayuda de un protocolo mediador.
- **Infraestructura**: Utilización de un medio como Internet, para realizar los procesos de Intercambio de documentos, metadatos y objetos digitales.
- **Estructura**: Uso de metalenguajes estructurados (XML, ASN1), como elementos que representan la semántica y sintaxis entre los sistemas de información.

La interoperabilidad es uno de los retos más grandes de las bibliotecas modernas, que buscan construir una biblioteca universal, donde las colecciones y los servicios son provistos por diferentes organizaciones con diferentes sistemas de cómputo, donde la información proviene de diferentes fuentes y es procesada de diferentes maneras y manejada por diferentes procesos de calidad. El propósito general que busca la interoperabilidad es el de construir una red de sitios académicos y bibliotecas, que permitirían la recuperación mundial dentro de una base de conocimiento heterogénea, independientemente de la localización física de los documentos proporcionados. "Los

[10] Wikipedia, http://es.wikipedia.org/wiki/Interoperabilidad
[11] Krsulovic , Ernesto. Blog de la Web Semántica. Octubre 19, 2002. http://www.dcc.uchile.cl/~ekrsulov/prj/ws-blog/

usuarios no deberían tener que navegar y buscar en varios servidores por separado. Deberían poder acceder a una sola interfaz que pueda conectar a todos los nodos diferentes de la red de los sitios que contienen la información y el conocimiento (bibliotecas)"[12].

Para ello, los sistemas interoperables deben implementar determinadas normas y estándares que permitan el acceso transparente y rápido a la información. Para lograrlo hay que utilizar y mezclar varios tipos de estas, cada una asociada a un tipo especial de problema: manejo de contenidos y su representación, su descripción, y los mecanismos de interconexión:

- Representación de contenidos (Objetos y Documentos digitales)
  - Formatos
  - Identificación única de documentos
  - Encapsulamiento de funcionalidad
  - Protección de Copyright
- Descripción de contenidos (Metadatos)
  - Niveles de descripción
  - Vocabularios controlados
  - Manejo de autoridades
- Mecanismos de interconexión (OAI, Z39.50, ZING SRU-SRW, Webservices, etc...)
  - Protocolos de comunicación
  - Formato de Mensajes, comandos y contenidos
  - Control de Errores
  - Duplicidad de la información

Como inconvenientes asociados a los anteriores ítems, podemos destacar el alto costo de adaptación de las normas y estándares en cualquier sistema de información y la constante evolución de los conceptos y sistemas que hacen que las normas y estándares cambien continuamente. Por tal motivo se debe valorar el coste de aceptación de estos

[12] Hilario Hernández, Las bibliotecas públicas en España, una realidad abierta. Edición electrónica. http://www.fundaciongsr.es/bp/index2.html

frente a la funcionalidad y hacer una proyección de la utilidad de estos en el mediano y largo plazo.

**Redes interoperables en la Web 2.0**

"La globalización no es solamente un fenómeno de integración de mercados, sino que debe favorecer de desarrollo de un conocimiento que no tenga restricciones ni fronteras, y que no se convierta en un pensamiento único que sacrifique los valores locales y las diferencias culturales, sino que debe convertirse en un puente hacia el desarrollo de las mismas".[13] Estamos entrando en la era de la Web 2.0; Una Web más inteligente, donde las maquinas trabajarán más en el manejo automático de la información, donde la Web semántica se impondrá sobre la Web sintáctica, donde la recuperación de información se realizará uniformemente desde cualquier sitio de Internet y no tendremos que desplazarnos de sistema en sistema para obtener lo que necesitamos. Este paradigma informacional lo estamos evidenciando desde hace algunos años con el incremento de bases de datos documentales (ó mejor bases documentales), que contrario a los modelos tradicionales de buscar sobre los metadatos exclusivamente, buscan sobre el documento en texto completo y además son capaces de realizar cruces y obtener relaciones entre múltiples documentos simultáneamente.

En este contexto los conceptos del Internet Invisible mencionados anteriormente dejan de ser validos, porque se supone (o se espera) que todos los sistemas de información documental brinden alternativas de interoperabilidad y permitan que sus contenidos sean accedidos directamente a través de su interfaz, y simultáneamente por cualquier sistema de búsqueda genérico o específico que exista en el mundo. Es claro evidenciar que estamos hablando principalmente de contenido académico, lo que constituye el verdadero capital intelectual de nuestra sociedad

Bajo este contexto vemos que REDINED al unificar y estandarizar su sistema de información para que sea interoperable y permita mejorar la recuperación de informacion, podría incrementar significativamente la visibilidad de cada uno de sus documentos, con el fin de que los autores de estos obtengan mayor reconocimiento nacional e internacional. Para citar tres ejemplos representativos que evidencian el

[13] Hno. Rodriguez Echeverría, Álvaro. La educación universitaria dentro de la misión Lasallista. Revista de la Universidad de La Salle. Vol. 42, Año XXVII, Julio-Diciembre de 2006.

poder de la interoperabilidad para la unificación de contenidos y su recuperación tenemos los siguientes buscadores:

- **Google Académico**[14]: Google Académico te permite buscar bibliografía especializada de una manera sencilla. Desde un solo sitio podrás realizar búsquedas en un gran número de disciplinas y fuentes como, por ejemplo, estudios revisados por especialistas, tesis, libros, resúmenes y artículos de fuentes como editoriales académicas, sociedades profesionales, depósitos de impresiones preliminares, universidades y otras organizaciones académicas. Google Académico ayuda a encontrar el material más relevante dentro del mundo de la investigación académica.
- **Scirus**: Buscador de Internet que permite buscar recursos en la red tanto páginas Web, como revistas electrónicas. A diferencia de otros buscadores, sólo recupera información de contenido científico.
- **Oaister**: Es un proyecto de la Biblioteca Digital de la Universidad de Michigan. Su meta es crear una colección de recursos digitales académicos de difícil acceso. Organiza el material localizado en las colecciones de más de 700 instituciones y ofrece cerca de un millón de registros de: libros electrónicos, artículos de revistas en línea, archivos de audio y video, imágenes, etc.

**REDINED como una red interoperable en la Web 2.0**

A continuación se describe el trabajo desarrollado en REDINED, cuyo objetivo principal consiste en dar mayor visibilidad de los contenidos existentes en la base de datos, y mediante el uso de la interoperabilidad con otros sistemas de información, unificar la recuperación y consulta de esta información desde cualquier motor de búsqueda (google, scirus, oaister, etc..), todo esto utilizando los principales protocolos y estándares internacionales. Para convertir el sistema información de REDINED en un sistema información Interoperable se dividió el trabajo de acuerdo los siguientes problemas: manejo de contenidos y su representación, su descripción, y los mecanismos de interconexión:

- **Representación de contenidos (Objetos y Documentos digitales)**

[14] Google. **Acerca de Google Académico.** http://scholar.google.es/intl/es/scholar/about.html <consultado el 04/09/2006>

Para representar cada registro la base de datos de REDINED como un objeto digital se deben tener en cuenta aparte de la descripción del documento, los diferentes formatos en que se encuentran los documentos en texto completo, un identificador único universal que permita diferenciar cada objeto digital de otro existente y un método de Protección de Copyright. Al analizar la base de datos de REDINED, según el manual de la base de datos global[15], encontramos que:

**Formatos:** Al realizar un análisis a la base de datos de REDINED, sobre los registros que poseen información de enlace a documentos en texto completo (campo 856, en los que aparecen 2407 enlaces), encontramos que predomina el uso de la extensión de archivos de PDF, el cual es usado en el 84. 34% de las veces frente a otros formatos. El formato PDF puede considerarse un estándar de facto para manejo de documentos digitales, lo que le permite asegurar su preservación y conservación a largo plazo frente a otros formatos de archivo. Es claro que el porcentaje documentos ofimáticos (Word, Excel, PowerPoint) enlazados apenas alcanza al 1% de la base de datos, los cuales se aconsejaría retirar y transformar a un formato más estandarizado. Asimismo se aconseja normalizar las extensiones de los documentos tipo página Web de las que encontramos variaciones entre documentos.htm y documentos.html.

| **Extensión de Archivo** | **Total** | **Porcentaje** |
|---|---|---|
| .doc | 10 | 0,42% |
| .htm | 106 | 4,40% |
| .pdf | 2030 | 84,34% |
| .pps | 3 | 0,12% |
| .ppt | 8 | 0,33% |
| .zip | 4 | 0,17% |
| .html | 200 | 8,31% |
| SIN ESPECIFICAR | 46 | 1,91% |
| **Total general** | **2407** | 100,00% |

Tabla 1. Extensiones de archivo en la base REDINED

[15] Calbet Roselló, Ernesto; Sánchez Palomar, Jorge. REDINED: Manual de la base de datos global – (investigación, innovación, recursos, analíticas). Noviembre de 2005. http://www.doredin.mec.es/documentos/rediman.pdf

**Identificación:** después de analizar la base de datos encontramos que el uso del primer campo (campo 1) de la base de datos es un excelente identificador único, que incluye dentro de sí datos como la unidad creadora, fecha, entre otros. Sin embargo para convertir este identificador existente en la base en un identificador estándar normalizado lo hemos de convertir en un identificador tipo URN (incorporando un esquema, un identificador dentro del repositorio y el campo antes mencionado), quedando así:

```
<scheme>oai</scheme>
<repositoryIdentifier>redined.mec.es</repositoryIdentifier>
<delimiter>:</delimiter>
<sampleIdentifier>oai:redined.mec.es:018200530125</sampleIdentifier>
```

**Protección de Copyright:** Respecto al manejo de los derechos de autor de los documentos publicados en la base de datos de REDINED, vemos que no hay un campo específico dentro la base de datos para manejar este tema. Sin embargo después de conversar con los creadores de esta base de datos, nos comentaron que: "Nosotros consideramos que todos los documentos depositados en **doredin.mec.es** son documentos financiados con fondos públicos y por tanto, que se pueden copiar indicando la fuente y no haciendo un uso comercial de los mismos. Las otras direcciones existentes en la base de datos, son enlaces de Internet que asumen su propia responsabilidad en el texto"[16].

- **Descripción de contenidos (Metadatos)**

Para mejorar el acceso a los documentos (para permitir su recuperación en entornos sobre los cuales no se puede buscar directamente en el documento), los bibliotecólogos han recurrido al método de realizar una descripción global de estos siguiendo metodologías específicas, y ofreciendo estas descripciones a todos los usuarios. Ahora bien debido al gran volumen de las fuentes y recursos que existen actualmente en Internet, se hizo necesario extrapolar los conocimientos existentes para establecer mecanismos que permitieran etiquetar, catalogar, describir y clasificar los recursos presentes en Internet, estos mecanismos son llamados comúnmente metadatos.

[16] Calbet, Ernesto. Correo electrónico recibido el día 20 de noviembre de 2006.

"Un metadato no es más que un dato estructurado sobre la información, o sea, información sobre información, o de forma más simple, datos sobre datos. Los metadatos en el contexto de la Web, son datos que se pueden guardar, intercambiar y procesar por medio del ordenador y que están estructurados de tal forma que permiten ayudar a la identificación, descripción, clasificación y localización del contenido de un documento o recurso Web y que, por tanto, también sirven para su recuperación"[17].

Al analizar el modelo de descripción utilizado por REDINED (metadatos), descubrimos que el modelo utilizado actualmente es una variante simplificada del estándar MARC 21, del cual se toman los principales campos y se modelan para ofrecer una mayor flexibilidad y facilidad en el momento de ingresar la información en Internet. Sin embargo el uso de este modelo específico no me permite estandarizar e intercambiar la información con otros sistemas de información, ya sea porque estos manejan el estándar MARC 21 de manera fidedigna ó manejan otros estándares de descripción. El primer paso para normalizar la información existente en REDINED, correspondió en realizar un mapeo de los campos encontrados en la base de datos y contrastarlos con los estándares de metadatos MARC 21 y DUBLIN CORE (ambos codificados en XML), generando los siguientes resultados:

| **REDINED** | **Dublin Core** | **MARC** | **Notas de Implementación** |
|---|---|---|---|
| 245^* | Title | 245$a | Incluye título propiamente dicho y título de serie |
| 440^*--440^d | | 440$a | |
| 100^*, | creador | 100$a, | Creador/Autor principal de la obra |
| 110^*, | | 110$a, | |
| 111^n—111^a | | 111$a, | |
| 653^* | Subject | 650$a, | Incluye los términos del Tesauro Europeo de educación |
| 655^* | | 650$a, | |
| 658^* | | 653$a | |

[17] María Jesús Lamarca Lapuente. Hipertexto, el nuevo concepto de documento en la cultura de la imagen. Tesis doctoral. Universidad Complutense de Madrid
http://www.hipertexto.info/documentos/indice.htm

| | | | |
|---|---|---|---|
| 500^* | Description | 500$a | Campo repetible en Dublín Core por cada subcampo del campo 520 de REDINED |
| 504^* | | 500$a | |
| ‘Objetivo:’520^a<br>‘Muestra’520^b<br>‘Proceso’520^c<br>‘Instrumento:’520^d<br>‘Tecnica:’520^e<br>‘Resultados’520^f<br>‘Conclusiones’520^g | | 520$a | |
| 700^*, | Contributor | 700$a, | Incorpora responsable y centro realizador de la obra |
| 710^a, | | 710$a, | |
| 711^a | | 711$a | |
| 260^a – 260^b | Publisher | 260$a$b | |
| 260^c | Date | 260$c | |
| 14^* | Type | Leader06, Leader07 | Descripción del tipo de material (libros, revistas, etc..) |
| 16^* | | 655$a | |
| right(v856^a,4)<br>300^* | Format | 856$q<br>300$a (para representar material fisico) | Para los documentos que contengan enlaces a documentos en texto completo. |
| 20^* | Identifier | 856$u | Utilizando notación URN |
| -- | Source | 786$o$t | URL del enlace al registro en la base REDINED |
| 18^* | Language | 41$a | |
| ‘Localizado en:’ v90^a – v90^f<br>‘publicado en:’ v773^a, -- v773^b, -- v773^c | Relation | 530$a | En MARC las relaciones se incorporan como notas |

| | | | |
|---|---|---|---|
| 11^* | Coverage | 651$a | |
| -- | Rights | 540$a | Se colocará por defecto un enlace URL que contenga un documento con los derechos de autor asociados al documento específico |

Tabla 2. Comparación de metadatos de REDINED, MARC21 y DUBLIN CORE

- **Mecanismos de interconexión (OAI-PMH, Z39.50, ZING SRU-SRW, Webservices, etc...)**

Los mecanismos de interconexión incluyen los protocolos de comunicación y formatos de intercambio de mensajes (**Cómo se recuperan/transportan**) entre los diferentes sistemas de información documental. El uso adecuado de estos, basados en la normatividad estandarizada existente, nos permitiría intercambiar automáticamente y transparentemente la información entre los diversos sistemas informáticos sin la intervención humana. Para lograr estos mecanismos, se debe manejar componentes de semántica asociados a unas reglas de mensajes o comandos.

Entre los principales mecanismos que me permiten realizar estos proceso se encuentran: Z39.50, OAI-PMH y ZING (SRU/SRW). Para el caso del sistema de información de REDINED se implementaron estos dos últimos, principalmente porque estos protocolos poseen características de última generación y son de fácil implantación. Además su funcionamiento está basado en la utilización del protocolo HTTP. Ambos permiten el intercambio de diferentes modelos de metadatos (MARC, Dublin Core, RFC 1708, etc..), y el uso del meta lenguaje XML para estructurar los documentos que se van a intercambiar. A continuación vamos a realizar una descripción de estos dos protocolos:

- **OAI-PMH (Protocolo de recolección de metadatos en la iniciativa de archivos abiertos)**

El protocolo de recolección de metadatos (OAI-PMH), es el mecanismo de trabajo que debe ser implementado en los sistemas de información documental para poder comunicarse entre si (interoperabilidad), este protocolo define todos los “comandos ó

verbos" necesarios para poder recolectar metadatos de forma automática y normalizada. Este protocolo define toda su funcionalidad en solo seis (6) verbos básicos, y algunos parámetros complementarios. En el caso de REDINED para utilizar este protocolo se debe invocar desde: **http://www.redined.mec.es/oai/oai.php**

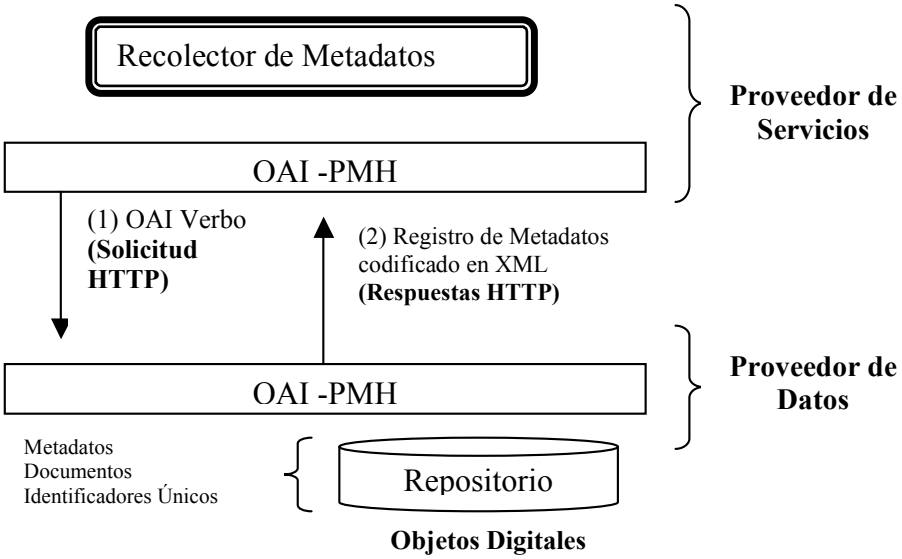


**Figura 2. Modelo de transacciones OAI-PMH**

La figura anterior[18] muestra el funcionamiento del protocolo OAI-PMH, básicamente este utiliza transacciones HTTP para emitir preguntas y obtener respuestas entre un proveedor de servicios y un proveedor de datos. Estas transacciones se realizan utilizando el método GET, que representa las operaciones de recepción y envío de información entre un programa cliente servidor a través del protocolo HTTP, estas constan de:

- Una dirección URL de un archivo que ejecuta un procedimiento
- Una lista de parámetros con la forma de pares del tipo: nombre_variable = valor.

Un ejemplo de una transacción HTTP es:
(http://www.redined.mec.es/oai/oai.php?verb=Identify)

[18] COLE, T., MISCHO, W. AND HABING, T. Introduction to the open archives initiative Protocol for metadata harvesting, introduction tutorial given at the ACM/IEEE Joint Conference on Digital Libraries, Houston, TX, (2003). (http://dli.grainger.uiuc.edu/publications/TWcole/)

Donde **http://www.redined.mec.es/oai/oai.php** indica la ubicación del archivo **oai.php** en el servidor de Internet de REDINED, cuando este archivo es invocado desde un navegador de Internet, se ejecutan unos procedimientos internos que son condicionados generalmente a unos parámetros asociados, el parámetro en este ejemplo es la palabra "**verb"**, cuyo valor es **"Identify"**. La tabla siguiente indica cuáles son los verbos que se pueden utilizar para extraer información del servidor de REDINED:

| **Verbo** | **Función** |
|---|---|
| **Metadatos acerca del repositorio** | |
| Identify | Descripción del repositorio |
| ListMetadataFormats | Lista de todos los tipos de metadatos soportados en el repositorio |
| ListSets | Lista de las particiones lógicas de información en el repositorio |
| **Verbos de recolección de metadatos** | |
| ListIdentifiers | Lista de los identificadores OAI únicos existentes en el repositorio |
| ListRecords | Lista de **n** registros de metadatos |
| GetRecord | Lista de un solo registro de metadatos |

**Tabla 3. Verbos del protocolo OAI-PMH**

- **ZING (Z39.50 Next Generation):** "Bajo este curioso nombre se engloba el último proyecto de la Z39.50 International Maintenance Agency que abarca una serie de iniciativas promovidas por el grupo de impulsores del protocolo Z39.50, mediante las que se pretende conseguir que tanto los contenidos intelectuales cómo semánticos del famoso protocolo Z39.50 sean más accesibles y de paso transformarlo en una herramienta más atractiva de cara a los nuevos proveedores de información, desarrolladores, distribuidores y usuarios ahorrando esfuerzos a la hora de ponerlo en marcha"[19] .

Esta iniciativa se compone básicamente de dos protocolos cuyo funcionamiento básico está regido por los mismos principios, estos protocolo son: SRW (Search and Retrieve Web Service) y SRU (Search and Retrieve URL). La diferencia de estos dos protocolos radica principalmente en el canal de comunicación utilizado entre los sistemas de información, sin embargo en su componente semántico y lógico son exactamente

[19] Fernández, Juan José; Garcia, Silvia. ZING Z39.50 International: Next Generation. http://www.absysnet.com/tema/tema25.html

iguales. En el caso de REDINED se implementó el protocolo SRU, para utilizar este protocolo se debe invocar desde:
**http://www.redined.mec.es/sru/sru.php**

Al igual que protocolo OAI, el protocolo SRU utiliza transacciones HTTP para emitir preguntas y obtener respuestas entre un proveedor de servicios y un proveedor de datos. Este es invocado por medio de unas operaciones o "verbos", los cuales permiten realizar una búsqueda estandarizada en el sistema de información documental. Este protocolo se define básicamente mediante la utilización de tres (3) verbos, todos ellos con algunos parámetros complementarios que aumentan la semántica del comando:

| **Verbo (Operación)** | **Función** |
|---|---|
| Explain | Operación que me brinda una descripción del sistema de búsqueda, y la descripción de los comandos a utilizar. |
| SearchRetrieve | Operación que me realiza una búsqueda en el sistema de información a partir de una sentencia dada en el lenguaje CQL (Common Query Language) |
| Scan | Operación que me permite conocer los términos de indización y por los cuales se puede recuperar algún componente o documento del sistema. |

Un ejemplo de una búsqueda utilizando el protocolo SRU es:
(http://www.redined.mec.es/sru/sru.php?version=1.1&operation=searchRetrieve&query=Educación&maximumRecords=10&recordSchema=dc )

Donde **http://www.redined.mec.es/sru/sru.php** indica la ubicación del archivo **sru.php** en el servidor de Internet de REDINED, cuando este archivo es invocado desde un navegador de Internet, se ejecutan unos procedimientos internos que son condicionados generalmente a unos parámetros asociados:

- **operation=searchRetrieve** (Comando principal que indica la acción principal a ejecutar)
- **version** = versión del protocolo.
- **query**=expresión de consulta (todos los resultados que tienen Educación en alguno de sus metadatos).
- **maximumRecords=<x>** cantidad de registros a visualizar por búsqueda.

- **recordSchema**=dc, se refiere al modelo de metadatos en XML para representar los documentos resultados de la búsqueda.

Vemos que el protocolo OAI está enfocado hacia la cosecha/recolección de metadatos distribuidos en varios sistemas de información (proveedor de datos) hacia un servidor centralizado donde se unifican toda las colecciones para su posterior búsqueda utilizando un lenguaje único (proveedor de servicios), mientras que el protocolo SRU está enfocado a realizar la búsqueda distribuida a través de los diversos sistemas de formación cuantas veces sea necesario (cada vez que se realice una búsqueda por parte de un usuario).

Es de señalar la importantísima sinergia que ha surgido entre el protocolo OAI-PMH y los protocolos de recuperación de información en bases de datos distribuidas dentro de ZING (Z39.50 International Next Generation) y denominados SRW/U (Search/RetrieveWeb Service). Esta correlación entre OAI y SRW/SRU era previsible puesto que en último término ambos protocolos tienen como objetivo facilitar la búsqueda y recuperación de la información, aunque afrontándola desde diferentes perspectivas. Por otra parte, se sustentan en la utilización de estructuras de metadatos específicas pero comunes entre ambos protocolos[20]. Podemos concluir que ambos modelos se complementan y trabajan en situaciones específicas dependiendo del modelo de metabúsqueda que se requiera.

## Conclusiones y trabajos futuros

Estamos evidenciando un momento en que todos los sistemas de información y bases de datos de existentes en el mundo utilizan o utilizarán algún protocolo de interoperabilidad (muy seguramente los mismos que se implementaron en la base de datos de REDINED). De esta forma podemos asegurar que cualquier persona podría ubicar documentos académicos con gran facilidad y acceder al texto completo de estos de forma inmediata. De lo contrario nos veremos abocados al problema de no encontrar

[20] Acuña, María José de; Agenjo, Xavier. "Los archivos en la era digital: el problema (y la solución) de los recursos electrónicos". En: El profesional de la información, 2005, noviembre–diciembre, v. 14, n. 6, pp. 407-413.

información relevante de calidad y seguiremos navegando en el mar infinito de documentos.

El desarrollo y aplicación de estos protocolos no requiere grandes conocimientos ingenieriles y de programación, simplemente se requiere conocer y aplicar los estándares y normas existentes para poder implementarlos en cualquier sistema de información. Actualmente en muchos de los sistemas para la administración de documental, bibliotecas y contenidos digitales y otros sistemas de información avanzados ya vienen incluidos estos estándares, y encontramos con que el problema principal consiste en que no lo sabemos utilizar.

Por ejemplo, para REDINED puede ser una gran oportunidad incorporar dentro de su sistema nuevas bases de datos en educación que se encuentren dispersas en el mundo, aquí podremos considerar a REDINED como un recolector y metabuscador de información, que al utilizar efectivamente protocolos de interoperabilidad, se conecte a otros sistemas de información distribuidos mundialmente, y permita extraer las referencias y los documentos en texto completo y que permitan a cualquier usuario buscar la información de manera centralizada utilizando únicamente la interfaz de REDINED (y porqué no, que integre dentro de su base de datos todo el sistema Eric y el sistema Reduc).

**Referencias**


Acuña, María José de; Agenjo, Xavier. "Los archivos en la era digital: el problema (y la solución) de los recursos electrónicos". En: El profesional de la información, 2005, noviembre–diciembre, v. 14, n. 6, pp. 407-413.

Bergman, M.K. (2001). The Deep Web: Surfacing Hidden Value. Journal of Electronic Publishing, 7(1).

Calbet Roselló, Ernesto; Sánchez Palomar, Jorge. REDINED: Manual de la base de datos global – (investigación, innovación, recursos, analíticas). Noviembre de 2005. http://www.doredin.mec.es/documentos/rediman.pdf

Colaboradores de Wikipedia. Controversia por la biografía de John Seigenthaler Sr. [en línea]. Wikipedia, La enciclopedia libre, 2006 [fecha de consulta: 3 de noviembre del 2006]. Disponible en <http://es.wikipedia.org/w/index.php?title=Controversia_por_la_biograf%C3%ADa_de_John_Seigenthaler_Sr.&oldid=5411696>.

COLE TIMOTHY W. Using OAI: innovations in the sharing of information. En: Library Hi Tech. Vol. 21 (2) . 2003. p. 115-117. (http://www.emeraldinsight.com/0737-8831.htm)

COLE, T., MISCHO, W. AND HABING, T. Introduction to the open archives initiative Protocol for metadata harvesting, introduction tutorial given at the ACM/IEEE Joint Conference on Digital Libraries, Houston, TX, (2003). (http://dli.grainger.uiuc.edu/publications/TWcole/)

Cornella, Alfons. Cómo sobrevivir a la infoxicación. Trascripción de la conferencia del acto de entrega de títulos de los programas de Formación de Posgrado del año académico 1999-2000. (Diciembre 2000)

Fernández, Juan José; Garcia, Silvia. ZING Z39.50 International: Next Generation. http://www.absysnet.com/tema/tema25.html

Fornas Carrasco, Ricardo (2002) Criterios para evaluar la calidad y fiabilidad de los contenidos en Internet. In Proceedings Contenidos y Aspectos Legales en la Sociedad de la Información (CALSI), Valencia (España).

FRANCIS PISANI Mejorar la Wikipedia y aprender a usarla. El Pais.Com. 05/01/2006. http://www.elpais.com/articulo/semana/Mejorar/Wikipedia/aprender/usarla/elpcibsem/20060105elpciblse_2/Tes/

Google. Acerca de Google Académico. http://scholar.google.es/intl/es/scholar/about.html <consultado el 04/09/2006>

Gómez Dueñas, Laureano Felipe, La Iniciativa de Archivos Abiertos (OAI), un nuevo paradigma en la comunicación científica y el intercambio de información. En Revista Códice, Universidad de La Salle. No 4, 2005.

Hilario Hernández, Las bibliotecas públicas en España, una realidad abierta. Edición electrónica. http://www.fundaciongsr.es/bp/index2.html

Hno. Rodriguez Echeverría, Álvaro. La educación universitaria dentro de la misión Lasallista. Revista de la Universidad de La Salle. Vol. 42, Año XXVII, Julio-Diciembre de 2006.

Krsulovic , Ernesto. Blog de la Web Semántica. Octubre 19, 2002. http://www.dcc.uchile.cl/~ekrsulov/prj/ws-blog/

María Jesús Lamarca Lapuente. Hipertexto, el nuevo concepto de documento en la cultura de la imagen. Tesis doctoral. Universidad Complutense de Madrid http://www.hipertexto.info/documentos/indice.htm

Netcraft. “November 2006 Web Server Survey”. Web Server Survey News. http://news.netcraft.com/archives/2006/11/01/november_2006_web_server_survey.html (Nov 1, 2006).